Baterakuntzan oinarritutako euskararen analizatzailea.

Oinarrizko PATR gramatika

Aldezabal I., Gojenola K., Sarasola K.

Ixa taldea

Informatika Fakultatea, UPV-EHU

{jibalroi, jipgogak, jipsagak}@si.ehu.es

1.1 Sarrera

Sintaxi konputazionala ez da atzoko kontua. 1950eko hamarkadan hasierako itzulpen-sistema primitibo haiek sortzen hasi zirenetik ikerlariek etengabe ekin diote hitzen arteko harremana eta perpausen egiturak konputazionalki lortzeari. Hasieran espero ez bezain zaila suertatu da, ordea, erronka hori. Gaur egunean, oraindik, ez dago testu errealetako edozein esaldi osorik analizatuko duen analizatzailerik. Ingeleserako sortu da ahalmen zabaleko sistemarik, baina sistema horiek ere oztopo itzel baten aurrean aurkitzen dira bere emaitzak aplikatu nahi dutenean: emaitza posible bakarra ez, baizik eta dozenaka-edo analisi posible lortzen baitituzte testu errealetako perpaus arruntak analizatzerakoan.

Gauzak horrela bi lerro nagusi bereizten dira gaur egunean sintaxi konputazionalean. Batetik, konputagailuaren laguntzaz hizkuntzaren deskribapen sakona lortzeko gero eta gramatika osatuagoak eraikitzea; bestetik, perpaus oso-osoen analisia lortzeko asmorik gabe, zatiak (ingelesez *chunk* esaten duten horietakoak: izen-sintagmak, adizlagunak, aditz-kateak, ...) bereizteko tresnak sortzea tresna horiek hizkuntza-teknologiako aplikazioak garatzeko baliagarriak izango direlako, esaterako, hizketa-sorkuntzan edo informazio-bilaketan.

Beste hizkuntzetarako zenbait gramatika eta analizatzaile sintaktikoren berri nahi izanez gero web-gune hauetan aurki daitezke baliabideak eta analizatzaileak:

- Natural Language Software Registry (http://registry.dfki.de)

- Computational Linguistics (On-line aurkezpenak)

  http://www.ifi.unizh.ch/CL/InteractiveTools.html#as-h2-3296

Edo hainbat sistemen artean ez galtzearren begiratu zuzenean honako hauek bakarrik:

- Ingeleserako analizatzaile sintaktiko bat: Conexor enpresaren web-gunean (ingelesaz gain, baita beste zazpi hizkuntzetarako ere).

  http://www.conexor.fi/testing.html#1

- Espainierarako bat: CliC. *http://www.ub.es/ling/labcat.htm*
- Frantseserako bat: Genevako Unibertsitatea. *http://latl.unige.ch*

Sintaxia konputagailuen bidez lantzeko lau lan-ildo nagusi sortu izan dira orain arte:

- Baterakuntzan oinarritutako testuingururik gabeko gramatikak (Shieber, 1986). Funtsean testuingururik gabeko gramatikak[1] dira, baina formalismo sintaktiko horri gehitzen zaizkio, batetik, osagai sintaktikoen informazioa biltzeko aukera ezaugarri-egituren[2] bidez, eta bestetik, erregela bakoitzari zenbait ekuazio definitu ahal izatea komunztadurak edo beste edozein murriztapen edo erlazio sintaktiko egiaztatzeko.
- Egoera finituko teknikak (Karttunen eta beste, 1997).
  Azaleko sintaxia baino ez dute lortzen formalismo hauek, hau da, perpausetako osagaiak bereiztea baino ez dute lortzen. Espresio erregularrak baliatzen dituzte osagaien eraketa definitzeko eta oso denbora txikian lortzen dituzte emaitzak.
- Murriztapen-Gramatikak (Karlsson eta beste, 1995). Hitz bakoitza modu isolatuan analizatzeko dauden aukera guztietatik abiatuz, murriztu egiten dira aukera horiek erregelen bidez hitz bakoitzak testuinguruan dituen beste hitzen arabera. Bukaeran, hitz bakoitzak irakurketa bakarra dauka, bertan bere kategoria, kasua, numeroa eta beste informazio sintaktikoak daudela, funtzio sintaktikoarekin batera. Formalismo hau, izatez, egoera finituko tekniken barruan kokatu beharko genuke, baina bere ezaugarri bereziak direla eta, aipamen berezia egin diogu.

---

[1] Ingelesez *Context Free Grammar* direnak.

[2] Ingelesez *feature structures* direnak.

- Metodo estatistikoak

  Lan-ildo honetan dabiltzan ikerlariek corpus handietatik informazio sintaktikoa (gramatikak, espresio erregularrak, ...) automatikoki deduzitzen saiatzen dira estatistika erabiliz, gero testu errealen gainean aplikatzeko. Normalean metodo estatistikoak lehenago aipatu ditugun hurbilpen horietako batekin batera aplikatzen dira, hau da, gramatika batekin elkarlanean aplikatzen dira metodo estatistikoak.

IXA taldearen ahaleginak lehenengo eta hirugarren ildoetatik joan dira orain arte euskararen sintaxia lantzeko orduan. Murriztapen-gramatiken inguruan egindako lanak Lafitteren omenezko biltzar honetako beste artikulu batean azaldu dira (Aduriz eta Arriola, 2001). Artikulu honetan testuingururik gabeko gramatiketan oinarritutako baterakuntza-formalismo baten gainean egindako lana aurkeztuko dugu.

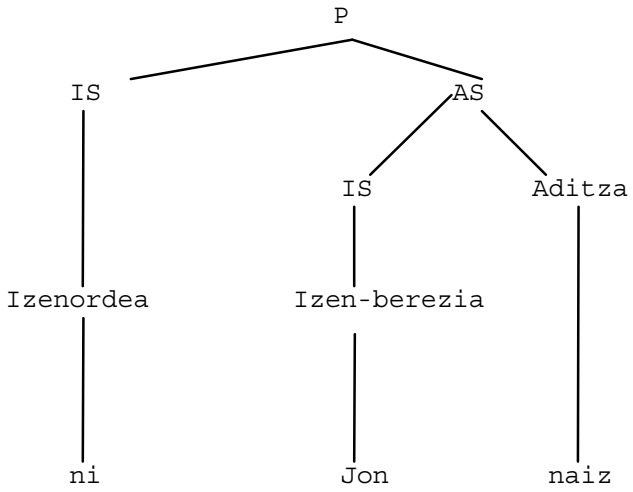
Hurrengo orrialdeetan honela egituratuko dugu artikulua: hasieran testuingururik gabeko gramatikak, baterakuntzan oinarritutakoak eta PATR gramatika-formalismoa labur aurkeztuko ditugu. Bigarren atalean, potoloena eta nagusia den atalean, PATR formalismoaren gainean sortu dugun gramatikaren ardatz nagusiak erakutsiko ditugu. Hirugarren atalean zenbait adibiderekin ilustratuko dugu gramatikarekin lortutako emaitzak zelakoak diren. Eta bukatzeko hiru atal labur erantsiko ditugu puntu hauek zehazteko: emaitzak zuhaitz gisa ikusteko tresna, euskararen hitz-mailako konplexutasuna dela-eta hitz solteak analizatzeko definitu behar izan dugun gramatika morfosintaktikoa eta, bukatzeko, azken atalean, *ondorioak* izenburupean gure ekarpenak laburtu eta etorkizunerako eginkizunak azalduko ditugu.

1.2 Baterakuntzan oinarritutako gramatika-formalismoak eta PATR

Baterakuntzan oinarritutako gramatikak testuingururik gabeko gramatiketan oinarritzen dira. Hauek Chomsky-k formalizatu zituen (Chomsky, 1957), eta modu honetako erregelen bidez adierazten dira:

Ingelesezko gramatika			*Euskarazko gramatika*		
S	→	NP	P	→	IS AS
VP	→	Verb NP VP	AS	→	IS Aditza
NP	→	Noun	IS	→	Izen-berezia
NP	→	Art Noun	IS	→	Izenordea

Erregelak '*a* → *b*' edo '*a* → *b c*' modukoak dira, non *a* ikurra ez terminala den eta *b, c* terminal edo ez terminalak. Ikur ez terminalak (S, NP, ...) kategoria sintaktikoak dira, eta ikur terminalak hiztegiko hitzak edo morfemak izango dira. Gramatikaren axioma edo hasierako ikurretik abiatuta (ingelesezko adibidean S (*sentence*) eta euskarazkoan P (perpausa)) eta erregelak erabiliz erator daitezkeen sinbolo terminalen kateak lengoaiarenak dira, besteak ez. Gehienetan lengoaiaren kateak zuhaitzen bidez adierazten dira. Adibidez, 1. irudian esaldi baten analisi-zuhaitza ikus dezakegu, aurreko taulan azaldutako euskarazko gramatikaren erregelak aplikatuz.



1. irudia. *'ni Jon naiz'* esaldiaren zuhaitz sintaktikoa.

Testuingururik gabeko gramatiken formalismoa sinplea da, baina arazoak daude lengoaia naturalaren fenomeno asko adierazi nahi direnean. Adibidez, aditza eta subjektuaren arteko komunztadura egiaztatu nahi badugu, numero, mugatasuna eta pertsonan, orduan 'AS → IS AS' erregelaren ordez beste erregela asko egin beharko genuke ('AS → IS_ergatibo_sing_3 AS_nor_nork_sing_3' eta honen moduko hainbeste erregela). Arazoa, beraz, adierazpen-ahalmen faltarena da. Hau gainditzeko, baterakuntzan oinarritutako formalismoak erabiltzen dira (Shieber, 1986). Hauen ideia nagusia testuingururik gabeko gramatiken osagai sintaktikoei informazioa gehitzea da, ezaugarri-egituren bidez, eta osagaien arteko erlazioak informazio horren gaineko ekuazioen bidez adieraztea. Euskararen kasuan, bere hitz-mailako fenomenoen aberastasuna eta tratatu beharreko egituren konplexutasuna aintzat hartuta, are eta beharrezkoagoa izango da baterakuntza erabiltzea.

Hau da Shieber-ek aurkezten duen erregela adibidea:

S → NP VP

S head = VP head

S head subject = NP head

Oinarrian testuingururik gabeko erregela bat dugu, perpausa (S) osatzeko modu bat adierazten duena, baina erregela horri ekuazio bi gehitu zaizkio osagaien arteko murriztapenak adierazteko. Lehenengoak dio esaldiaren burua edo gunea (head) eta aditz-sintagmarena gauza bera direla, eta bigarrenak esaldiaren subjektua aditz-sintagmaren aurretik azaltzen den izen-sintagma hori dela. Ekuazio horiek aplikatuz ezaugarri-egitura bat izango da esaldiaren informazio sintaktikoa. Hona hemen adibide partzial bat:

cat: S							
head:	form: finite						
	subject:		agreement:	number: sing			
				person: 3			

Baterakuntzan oinarritutako zenbait formalismo garatu dira han eta hemen, horien artean PATR, LFG (euskararen sintaxiaren formalizaziorako erabilia Abaitua (1988) eta Zubizarreta-ren (1992) lanetan), GPSG (Gazdar eta beste, 1985) eta HPSG (Pollard eta Sag, 1994) aipa ditzakegu. Gure kasuan, formalismo hauetatik guztietatik PATR formalismoa aukeratu genuen euskararen deskribapena egiteko, bi arrazoi nagusi kontuan izanda: lehenengo, analizatzaile konputazionala egiteko oinarrian dugun EDBL datu-base lexikala erabili beharko dugu (Aduriz eta beste, 1998), eta datu-base honek ez du LFG edo HPSG moduko formalismo batek beharko lukeen informazio konplexu osoa. Bigarren, PATR formalismoa malgua eta sinplea da, eta honengatik errazagoa izan da euskararen lehen tratamendua egiteko, ondorengo lanetarako HPSG edo LFGren inplementazioak baztertu gabe. Esan dugunez, testu errealak tratatzeko analizatzaile bat eraikitzea izan da gure helburua eta, beraz, (Abaitua, 1988) bezalako lanetan landu den gramatika konputazional sakona baino, nahiago izan dugu maizen gertatzen diren osagai sintaktikoak ezagutzea, honetarako PATR formalismo egokia izanik.

1. taulako gramatika txikiaren gainean azalduko ditugu PATR formalismoaren ezaugarri nagusiak.

R1. X0 → X1 X2
  X0 kat = AS
  X1 kat = IS
  X2 kat = AS
  X1 kas = erg
  X2 azpikat erg kom = X1 kom
  X0 = X2
	R2. X0 → X1 X2
	  X0 kat = AS
	  X1 kat = AS
	  X2 kat = IS
	  X2 kas = erg
	  X1 azpikat erg kom = X2 kom
	  X0 = X1

R3. X0 → X1 X2
  X0 kat = IS
  X1 kat = ize
  X2 kat = knmdek
  X1 azp = arr
  X0 gune = X1
  X0 kas = X2 kas
  X0 kom = X2 kom
	R4. X0 → X1
	  X0 kat = AS
	  X1 kat = adt
	  X0 azpikat = X1 azpikat

1. taula. Euskararen gramatika baten adibidea PATR formalismoan.

Lehen lerroko ezkerreko erregela (R1) aditz-sintagma bat (AS) eta izen-sintagma bat lotzeko erregela da. X0 osagaia (gurasoa) X1 eta X2 osagaiak konbinatuz lortzen da. Testuingururik gabeko erregela baten bidez ‘AS → IS AS’ adieraziko genuke. Ekuazioek osagaien arteko murriztapenak adierazteko eta osagai sintaktikoak nola osatu behar diren zehazteko erabiltzen dira. Horrela, lehen hiru ekuazioek osagaien kategoriak zeintzuk izan behar diren zehazten du, erregela aplikatu ahal izateko. ‘X1 kas = erg’

laugarren ekuazioak izen-sintagma hori kasu ergatiboan egotea eskatzen du. Bostgarren murriztapenak aditz-sintagma eta izen-sintagmaren arteko komunztadura (numero, mugatasuna eta pertsona) egiaztatzeko balio du. Seigarrenak ('X0 = X2') dio gurasoa aditzaren proiekzioa dela eta, beraz, beraien informazioak berdinak direla. R2 erregelak gauza bera adierazten du, baina osagaien ordena aldatuz, hau da, 'AS → IS AS'. Modu honetan euskararen ordena librea tratatzen da. Antzeko erregelak definitu beharko dira izen-sintagma absolutibo eta datiboentzat, baita adizlagun eta mendeko esaldientzat.

1. taulako bigarren lerroan izen-sintagma osatzeko R3 erregela agertzen da ('IS → ize knmdek'). Horrela, izena eta kasu eta numeroaren marka daukan atzizkia lotuko dira. R4 erregelak dio kasurik sinpleenean aditz trinko bat aditz-sintagma dela. Hasierako aditz-sintagma honetatik abiatuta izen-sintagmak eta adizlagunak lotuko zaizkio ezker eta eskuinetik, esaldi konplexuak sortuz.

L1. X0 sar = dakar X0 kat = adt X0 azpikat erg kom num = hu X0 azpikat abs kom num = hu X0 err = ekarri	L2. X0 sar = dakarte X0 kat = adt X0 azpikat erg kom num = hk X0 azpikat abs kom num = hu X0 err = ekarri
L3. X0 sar = ak X0 kat = knmdek X0 kas = abs X0 kom num = hk X0 kom mug = m	L4. X0 sar = ek X0 kat = knmdek X0 kas = erg X0 kom num = hk X0 kom mug = m
L5. X0 sar = gizon X0 kat = ize X0 azp = arr	L6. X0 sar = txakur X0 kat = ize X0 azp = arr

2. taula. Euskararen lexikoi baten adibidea PATR formalismoan.

2. taulan lexikoi baten adibidea dugu. Lehen bi sarrerak (L1 eta L2) aditzak adierazteko dira. Aditz bakoitzeko bere kategoria (*adt,* aditz trinkoena) eta komunztadurari buruzko informazioa agertzen da. L1 sarreran, adibidez, *dakar* aditza nor-nork motakoa dela adierazten da eta, gainera, *ergatibo eta absolutibo* diren osagaiak singularreko hirugarren pertsonan (hu) doazela. L2 sarreran, *dakarte*-ren kasuan, *ergatibo* den osagaia pluraleko hirugarren pertsonan (hk) doala adierazten da. L3 eta L4 sarrerak kasu-marka atzizkiak deskribatzeko dira, adibide horietan ergatibo-singularra (*-ak*) eta ergatibo plurala (*-ek*). Azken lerroan bi izen arrunt agertzen dira.

Lexikoi hori eta 1. taulako gramatika hartuta, analizatzaile sintaktiko batek erabaki dezake '*gizonek dakarte*' edo '*dakar txakurrak*' esaldi zuzenak direla eta, aldiz, '*gizonek*

dakar' bezalakoak okerrak, ez baitu betetzen komunztadura egotea eskatzen duen murriztapena ('`X2 azpikat erg kom = X1 kom`' 1. taulako R1 erregelako murriztapena).

PATR formalismoaren adibide hori ikusita, hurrengo atalean euskararako landu dugun gramatikaren azalpena emango dugu.

1.3 Gramatikaren deskribapena

Euskararen sintaxiaren tratamendua egiteko, hauek dira kontuan izan ditugun aspektu nagusiak:

- Morfema izan da analisiaren oinarrizko unitatea deskribapen gehienetan, (Goenaga, 1980; Abaitua, 1988), eta berdin euskararen antzeko hizkuntzentzat ere. Honek esan nahi du bai morfologia bai sintaxia, egitura sintagmatiko beraren osagaiak kontsidera daitezkeela, beraien arteko muga argirik gabe. Adibidez, '*gizon* + *handi* + *-a*' sintagman, *-a* morfema sintakoki ez da lotzen adjektiboarekin, baizik eta izen-multzoarekin; horrela deskribapen gramatikala orokorrago eta sinpleagoa eginez.

- Informazio aberatsa dago lexikoian. Osagai lexikal bakoitzak (eta osagai horiek oinarritzat hartuz sortutako osagai sintaktikoak ere), informazio desberdina dauka: kasua, mugatasuna, funtzio sintaktikoa, ... Informazio hori guztia konbinatzea izango da gramatika sintaktikoaren eginkizun nagusia.

- Aditzaren azpikategorizazioa oraindik landu gabe dago. Aditza elementurik funtsezkoena izango da sintaxiaren deskribapenean, teoria sintaktikoetan zein sistema aplikatuetan. Aditzaren informazioan azpikategorizazioarena da konplexuena, aditz bakoitza zein motatako osagaiekin konbinatzen den zehazten duena. Euskararen kasuan, nahiz eta aditz laguntzaileak informazio asko eman (subjektu, objektu zuzena eta zeharkako objektuaren kasua, numeroa, eta pertsona), oraindik EDBL datu-base lexikalean aditz nagusi bakoitzaren informazio propiorik ez dago.

- Aditzaren komunztadura subjektu, objektu zuzena eta objektu ez zuzenarekin. Euskaraz komunztadura dago aditza eta ergatibo, absolutibo eta datibo kasuko osagaien artean.

- Perpaus-mailako osagai sintagmatikoen ordena librea. Jakina da euskaraz perpausaren osagai nagusien ordena libre samarra dela. Hau da, subjektua,

objektua, adizlaguna eta aditza emanda, beraien permutazio guztiak (hogeita lau) dira posible:

```
Txakurrak  egunkaria  ahoan  zekarren.
  subj       obj       adlg   aditza
```

Esan behar da ere malgutasun hori perpaus-mailan bakarrik ematen dela, beste osagaietan (izen-sintagma edo mendeko perpausak adibidez) askoz mugatuago dagoelako.

Gramatikaren azalpenarekin hasteko, ikus dezagun lehenago zelako izen-sintagma edota adizlagun onartzen dituen gure gramatikak, eta geroago aztertuko dugu perpaus-mailakoa. Izen-sintagmaren mailan hiru eraketa posible bereizi ditugu:

1. Buru gisa izen arrunt bat dakartenak. Azkeneko hitzean deklinabide-atzizkia etortzen da beti (*knmdek* esaten diogu horri). Izenaren aurretik izenlagun bat (*izlg*) edota determinatzaile bat (*det*) aukerazkoak dira. Atzetik adjektibo bat (*adj*) edota determinatzaile bat (*det*) aukerazkoak dira. Ekuazioen bidez kontrolatzen da zer determinatzaile etor daitekeen aurretik eta zein atzetik.

(izlg) +	**(det) +**	**ize +**	**(adj) +**	**(det) +**	**knmdek**
etxeko		*altzari*	*zahar*	*hori*	*ekin*
etxeko	*lau*	*altzari*	*zahar*		*etan*
etxeko		*altzari*	*zahar*		*ari buruz*

2. Buru gisa izen berezi bat dakartenak. Izen bereziaren aurretik izenlagun bat aukerazkoa da, baina ez da onartzen determinatzailerik, ezta adjektiborik ere eraketa mota honetan.

(izlg) +	**izb +**	**knmdek**
Donostiako	*Peru*	*ri*

3. Buru gisa izenorde bat dakartenak. Deklinabide atzizkia baino ez da onartzen kasu onetan.

ior +	**knmdek**
ni	ri

Horrelako egiturak onartu ahal izateko, osaketa sinpleenetik hasi eta konplexuenera heltzeko, *is1*, *is2* eta *is3* kategoria laguntzaileak bereizten ditugu. Osaketa konplexuena biltzen duen *is3* motako osagai bati deklinabide-atzizkia (*knmdek*) lotuz *isk* lortzen dugu. Definitzen ditugun *isk* horiek ez dira izen-sintagmak bakarrik, izen-multzo bati edozein deklinabide-atzizki erantsita lortzen diren guztiak baizik. Beraz, *isk* horiek batzuetan izen-sintagma eta beste batzuetan adizlagun ditugu, beraien arteko diferentzia nagusia kasua delarik.

Deklinabide-atzizkia modu orokorrean hartzen dugu. Kasu, numero eta mugatasunaren informazioak dakartzan atzizki multzoari *knmdek* esaten diogu. Are gehiago, posposizio kasuetan (adibidez *-ari buruz* edo antzekoak erabiltzen ditugunean) atzizkia bera gehi beste hitz bat ere hartzen du *knmdek* delako horrek.

Analizatzen diren izen-sintagmen osaketa ondoko erregela[3] hauetan ikus daiteke:

is1 →	*ize adj*	*etxe EDER*
|	*ize*	*etxe*
is2 →	*det is1*	*ZENBAIT etxe eder*
|	*is1 det*	*etxe eder BAT*
|	*is1*	*etxe*
|	*izb*	*JON*
is3 →	*izlg is2*	*MENDI HORRETAKO zenbait etxe eder*
|	*is2*	*zenbait etxe eder*
|	*ior*	*ZU*
isk →	*is3 knmdek*	*etxe ederrEKIN*
		mendiko zenbait etxe ederrAK
		mendiko zenbait etxe RI BURUZ

Izen-sintagmaren deskribapena bukatzeko esan dezagun izenlagunaren egitura nagusiak *isk* osagaien egitura bera duela, baina kasua genitiboetako bat izan beharko dela:

---

[3] Benetako erregelen sinplifikazioa erakusten dugu hemen. Hauek ez dira gure gramatikako erregelak osotasunean, horiek murriztapenak adierazteko eta osagai sintaktiko berriak sortzeko ekuazio-multzoa izango baitute.

izlg → is3 + knmdek(gen/gel) *mendi horretaKO*

izlg → as + erlt *nik ikusi duDAN*

Perpausa analizatzeko orduan subjektua ez dugu bereiziko eta horrela perpausa aditz-sintagma gisa hartuko dugu beti. Aditz-sintagma sinpleena aditza besterik ez duena da; aditz trinkoa zein aditz erroa gehi laguntzailea, aukera biak onartzen dira. Aditza ezagutu ondoren aditzaren ezker zein eskuinaldean ager daitezkeen osagaiak banan-banan onartuko dira, ondoan azalduko ditugun erregelak erabiliz:

1. Kasu nuklearrak analizatzeko erregelak. Ergatibo, absolutibo edo datibo kasuak hartzeko erregela hauetan ekuazioen bidez numero, kasu eta pertsona ezaugarrien komunztadura egiaztatzen da. Horrela, esate baterako, '**Peruk txakurrak ekarri du*' bezalako esaldiak ez dira onartuko. Erregelak bikoiztuta daude kasu horiek aditzaren aurretik zein atzetik onartu ahal izateko:

   as→ isk(erg) as GIZONEK ikusi dute

   as→ isk(abs) as GIZONAK ikusi dituzte

   as→ isk(dat) as GIZONARI eman dio

   as → as isk(erg) ikusi dute GIZONEK

   as→ as isk(abs) ikusi dituzte GIZONAK

   as→ as isk(dat) eman dio GIZONARI

2. Adjuntuak tratatzeko erregelak. Ergatibo, absolutibo edo datibo ez diren kasuak hartzeko erregelak dira hauek. Hemen eta hurrengo puntuetan ez ditugu erregelak era bikoiztuan erakutsiko, baina suposatu behar da elementua aditzaren atzetik ere onartzeko beste erregela bat definitu dela:

   as→ isk(ez da abs, erg edo, dat) as GIZON HORREKIN ikusi dute

3. Adberbioak tratatzeko erregelak:

   as→ adb as GAUR egin dut

4. Mendeko perpausak tratatzeko erregelak: konpletiboak, zehargalderak, moduzkoak eta denborazkoak:

   as → mend-modu-denb as HONA NENTORRELA ikusi dut

   as → mend-zehargaldera as EA JOAN DEN galdetu du

as → mend-konp as ETORRI DIRELA jakin da

Guztira 90 erregela definitu dira eta batez beste 15 ekuazio ditu erregela bakoitzak. Esan bezala erregelak hemen aurkeztu baino konplexuagoak dira, bai notazio aldetik, bai ekuazioen aldetik, bai ñabarduren aldetik. Erregela baten (*is3 → izlg + is2*) benetako itxura ikus daiteke hemen:

```
X0 ---> X1, X2
        X1 kat              = izlg
        X2 kat              = is2
        X0 kat              = is3
        X0 sint kom         = X2 sint kom
        X0 gunelex          = X2 gunelex
        X0 sint osgk izlg   = X1 sint izlg
        X0 sint osgk adj    = X2 sint osgk adj
        X0 sint osgk detaur = X2 sint osgk detaur
        X0 sint osgk detatz = X2 sint osgk detatz
        X0 sint nag kom     = X2 sint nag kom
        edo[eta[ X2 gunelex nag kat = eli,
                 X0 forma = X1 forma],
                 X0 forma = $($(X1 forma, "+"),
                              X2 forma)]
```

Zer analiza dezake gramatika honek? Zein da bere estaldura? Maila lexikoan oso estaldura handia dagoenez (EDBL datu-base lexikaleko 70.000 sarrerak erabiltzen dira analisian) esan dezakegu testu errealetako ia izen-sintagma eta adizlagun guztiak analiza daitezkeela. Berdin esan dezakegu esaldi sinpleen kasuan, hau da, esaldian puntuazio-markarik gabe honelako elementuen sekuentzia agertzen bada:

- Aditza
- Kasu nuklearrak (ergatibo, absolutibo eta datiboa)
- Adjuntuak
- Adberbioak
- Nominalizazioak
- Erlatibozko menpeko perpausak
- Konpletibo menpeko perpausak
- Moduzko menpeko perpausak
- Denborazko menpeko perpausak
- Zehargalderak

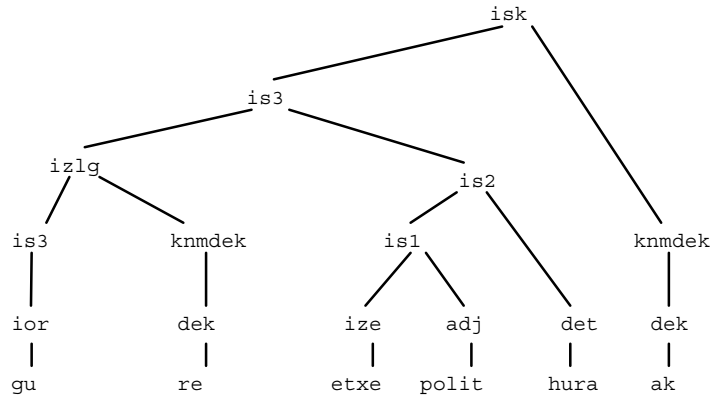
Adibide erreal gisa, hauxe da euskara-ikasle batek idatzitako corpus batetik atera dugun esaldi bat gramatikak osorik analizatzen duena[4]: '*Jakina da gaurko gizonak daraman bizimoduak ez duela antzarik antzinako gizonek zeramatzatenekin'*.

---

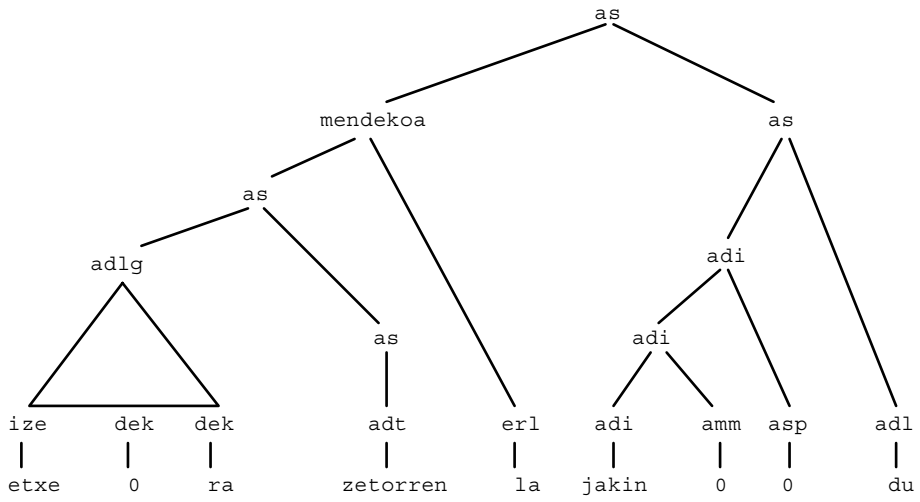
4 Analizatzaileak onartu du '*antzarik*' hitz okerra. Hori gertatu da hitzak aztertzean morfologiako errore tipikoak tratatzeko gai delako.

1.4 Adibideak

2. irudian *'gure etxe polit hark'* izen-sintagmari dagokion analisia ikus daiteke. Hor *'gure'* izenlaguna eta *'etxe polit hura' is2* bilduz *is3* motako osagai bat lortu da, eta berau *'ak'* deklinabide-atzizkiarekin batuz lortu du gramatikak hitz-kate osoa *isk* gisa ezagutzea. 3. irudian *'etxera zetorrela jakin du'* aditz-sintagmari dagokion analisia ikus daiteke. *'etxera zetorrela'* mendeko perpaus konpletibo gisa lotzen zaio esaldi nagusiari.



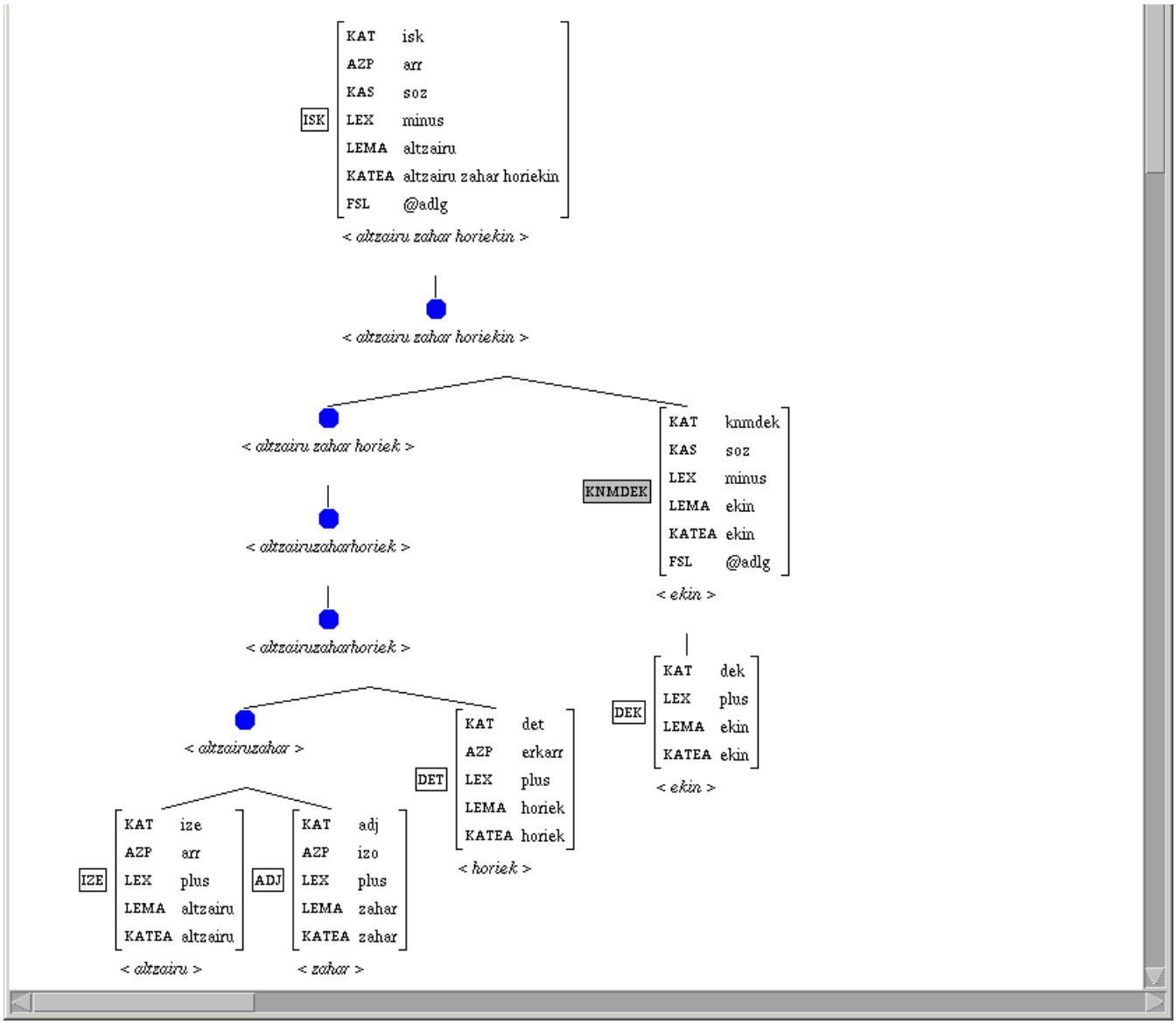
2. irudia. *'gure etxe polit hark'* izen-sintagmaren zuhaitz sintaktikoa.



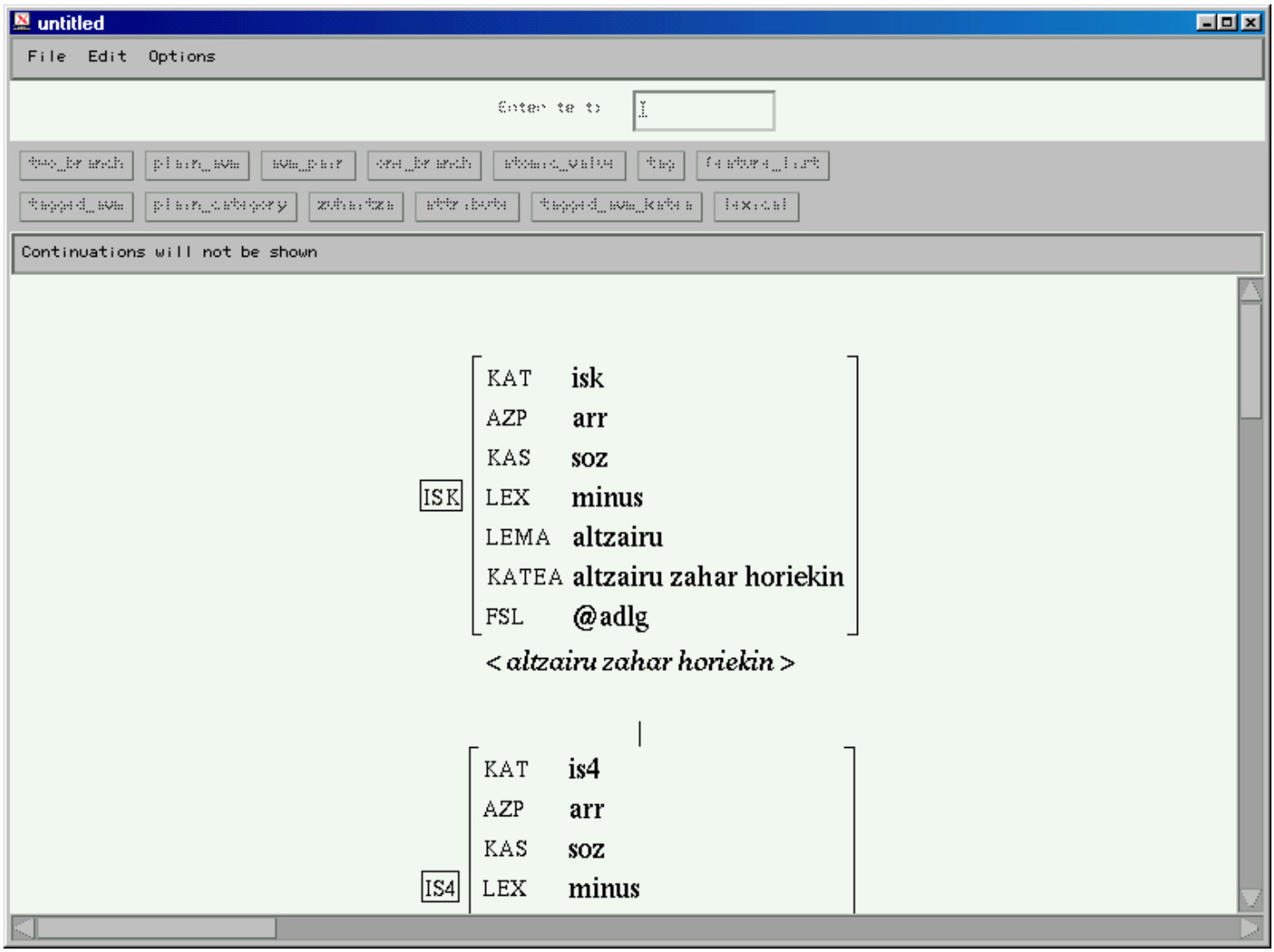
3. irudia. *'etxera zetorrela jakin du'* esaldiaren zuhaitz sintaktikoa.

1.5 Zuhaitz sintaktikoak ikusteko tresna

Erakutsi ditugun adibideetako esaldiak oso laburrak izan dira eta hala ere pantailan testu moduan lortzen ditugun emaitzak luzeak dira. Testu errealetako esaldiak analizatzean lortzen diren emaitzak oso-oso luzeak dira, alde batetik esaldiak luzeak direlako, eta bestetik osagai sintaktiko bakoitzak informazio aberatsa duelako. Horrelakoetan ez da batere erosoa emaitzen azterketa. Emaitzak zuhaitz gisa bistaratzeko tresna bat garatzen ari gara arazo horri ekiteko (Lopez de Lacalle, 2001). Sistema horretan posible da zuhaitza ikusteko leihoa ezker-eskuinera eta goitik behera mugitzea eta gainera aukerak gehitu dira nodoetako informazioak osorik edo partzialki erakusteko. 4. irudian ikus daiteke 'altzairu zahar horiekin' adizlagunaren emaitza. Bertan nodo batzuetan



4. irudia. *'altzairu zahar horiekin'* adizlagunaren zuhaitz sintaktikoa.



5. irudia. 4. irudiko zuhaitzaren zati bat erakusteko leihoa.

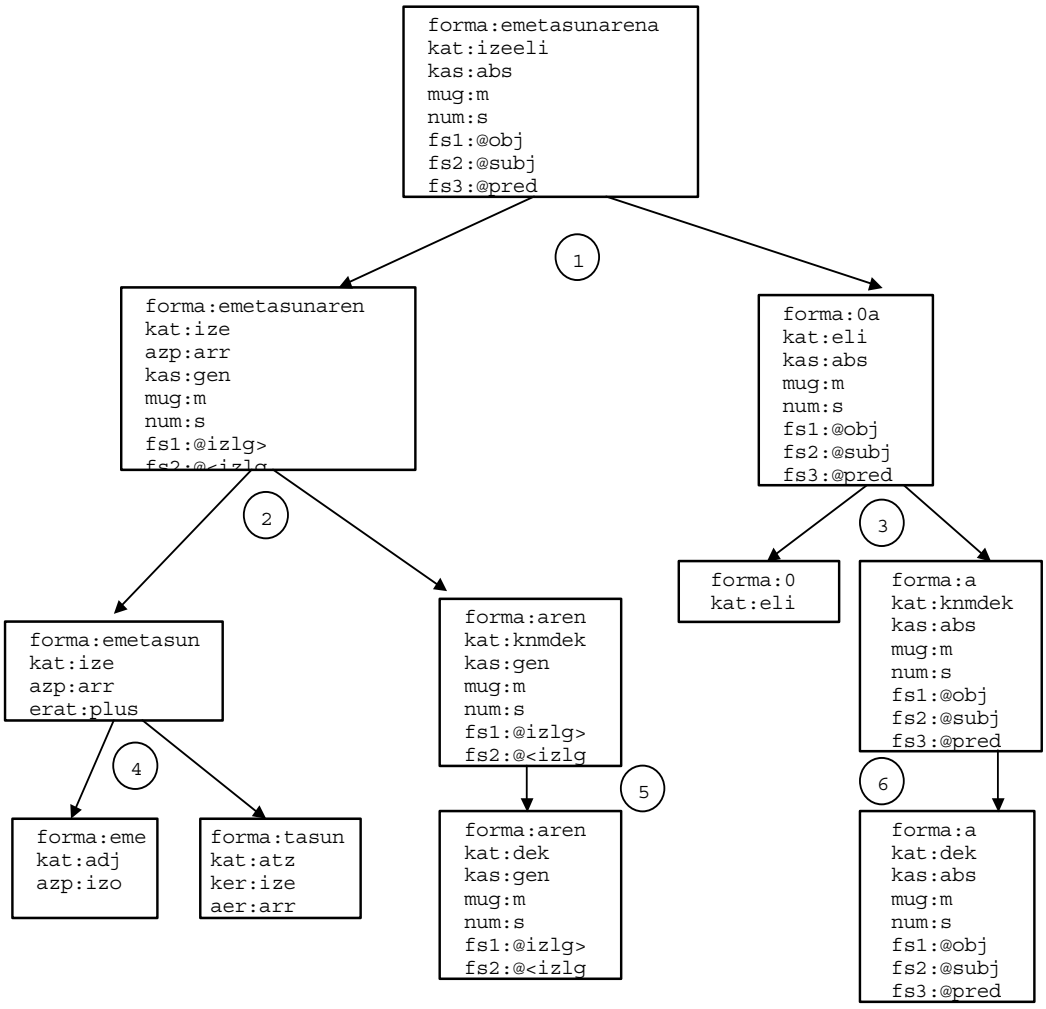
ezaugarriak erakusten dira eta beste batzuetan ez, azken hauetan informazio-ezkutatze hori puntu lodi batez adieraziz. 4. irudiko pantaila horretan azpizuhaitz bat aukera daiteke eta tamaina handiagoan ikusi gero 5. irudiko pantailan ikus daitekeen bezala. Modu horretan zuhaitz handietan nabigatu ahal izango da, zuhaitzaren adarrak ezkutatu edo informazio sintaktikoak handituz. Gure epe laburreko helburua izango da tresna honen bidez esaldien analisiak aztertu eta corpusen azterketa egin ahal izatea.

1.6 Hitz mailako gramatika morfosintaktikoa

Gramatika sintaktiko bi garatu ditugu. Bat, orain arte artikulu honetan deskribatu duguna, esaldiak analizatzeko balio duena, eta bestea hitz-mailako analisi morfosintaktikoa lortzeko erabiltzen dena. Beste hainbat hizkuntzatan ez, baina

euskaraz bigarren bertsio hori beharrezkoa da hainbat prozesaketa egin ahal izateko, hala nola, hitzen kategoria-etiketatze automatikoa edo desanbiguazioa, lan hauetarako definitu diren formalismo askotan (hitzen desanbiguazio estatistikoa edo murriztapen-gramatiketan egiten den bezala) hitza delako analisirako unitatea. Kontua da euskaraz hitz solteak analizatzean ere esaldi mailako egiturak errepresentatu behar izaten ditugula. Nola errepresentatu bestela '*emetasunarena*' edo '*zekarrenarekin*' bezalako hitzak? Ikus 6. irudian nolako analisi-zuhaitza lortzen den '*emetasunarena*' hitza analizatzeko. Zuhaitz horretan agertzen da hitz horren azpian elidituta dagoen osagaia ('*emetasunaren –X-a*' sintagmaren parekoa).

Beraz, hitz isolatuen analisi sakona lortzeko ere gramatika oso bat definitu behar izan dugu (Aduriz eta beste, 1999, 2000). Noski, esaldi mailakoa baino sinpleagoa da, eta hainbat erregela oso antzekoak dira gramatika bietan. Esaldi mailakoan 90 erregela definitu dira, eta hitz-mailakoan, berriz, 45 erregela nahiko izan dira.



6. irudia. *'emetasunarena'* hitz-formaren hitz-mailako analisia.

1.7 Ondorioak

Lan honetan euskarazko analizatzaile sintaktiko konputazionala aurkeztu dugu. Maila lexikoan oso estaldura handia duenez (EDBLko datu base lexikaleko 70.000 sarrerak erabiltzen dira analisian) esan dezakegu testu-corpusetatik (egunkariak edo testu idatziak) izen-sintagmak eta adizlagunak tratatzeko estaldura ia osoa dela, eta esaldi sinpleak analizatzen ere egokia dela.

Espero dugu tresna hau hizkuntzalarientzat baliagarria izango dela etorkizuneko euskararen gramatika oso baten hazi gisa, edo beste azterketetarako corpusak eta adibideak biltzeko. Alde konputazionalago batetik, analizatzailea erabilgarria da zenbait aplikaziotarako: informazioaren erauzketa edo desanbiguazioaren aurretik hitzen eta sintagmen analisiak osatzeko. Aplikazio hauek garatzeko, gure helburua tresna hauek hizkuntzalaritza konputazionalean lan egiten ari den komunitatearen eskuetan jartzea izango da.

Gramatika osatzeko bidean, eta garrantzi handiko hurrengo pauso gisa, aditzen azpikategorizazioari buruzko informazioa behar dugu esaldi-mailako analisi sakonak egin ahal izateko. Orain arte egindako gramatika hau erabiltzen ari gara aditzen informazio hori automatikoki lortzeko, testu idatzietako informazioa analizatuz, eta emaitza onak lortu dira (Aldezabal eta beste, 2001), nahiz eta oraindik jorratzeko bide luzea dugun.

1.8 Bibliografia

Abaitua, J. 1988. *Complex predicates in Basque: from lexical forms to functional structures*. Doktoretza-tesia, University of Manchester.

Abaitua J., Aduriz I., Agirre E., Alegria I., Arregi X., Artola X., Arriola J.M., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Sarasola K., Urkia M., Zubizarreta J.R. 1992. *Estudio comparativo de diferentes formalismos sintacticos para su aplicacion al euskara*. Barne-txostena, UPV/EHU/LSI.

Aduriz I., Agirre E., Aldezabal I., Arregi X., Arriola J.M., ArtolaX., Gojenola K., Maritxalar A., Sarasola K., Urkia M. 1999. *MORFEUS: Euskararako analizatzaile morfosintaktikoa*. Barne-txostena, UPV/EHU/LSI/TR 1-99.

Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A., Insausti J. M. 1998. *EDBL: a Multi-Purposed Lexical Support for the Treatment of Basque*. Proceedings of the First International Conference on Language Resources and Evaluation.Vol II. pp 821-826. Granada. Maiatza 28-30.

Aduriz I. 2000. *Morfologiatik Sintaxira Murriztapen Gramatika baliatuz*. Tesi-txostena, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Aduriz I., Arriola J. M. 2001. *Euskararen murriztapen gramatika. Desanbiguazio morfologikoaren tratamendua, azterketa sintaktikoaren lehen urratsak eta aplikazioak.* P. Lafitteren sortzearen mendemuga, Euskaltzaindia (Gramatika batzordea), Baiona.

Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Sarasola K., Goenaga P. 2001. *Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus*. Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural, Jaén.

Arriola J.M. 2000. *Hauta-Lanerako Euskal Hiztegi-ko informazio lexikalaren erauzketa erdi-automatikoa eta bere integrazioa sistema konputazional batean*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Chomsky, N. 1957. *Syntactic structures*. The Hague: Mouton.

Gazdar G., Klein E., Pullum G., Sag I. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press.

Gojenola K. 2000 *Euskararen sintaxi konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan*. Doktorego-tesia. UPV-EHUko Informatika Fakultatea.

Goenaga P. 1980. *Gramatika bideetan*. Erein

Lopez de Lacalle J. 2001. *Euskararen analizatzaile sintaktikoaren intefazea*. Karrera-bukaerarako proiektua. Euskal Herriko Unibertsitatea, Lengoaia eta Sistema Informatikoak.

Karlsson F., Voutilainen A., Heikkila J., Anttila A. 1995. *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Karttunen L., Chanod J-P., Grefenstette G., Schiller A. 1997. *Regular Expressions For Language Engineering*. Natural Language Engineering.

Pollard C., Sag I. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press.

Shieber S.M. 1986. *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes, 4 zenbakia, Stanford.

Zubizarreta J.R. 1992. *Un modelo funcional de diálogo para diálogos orientados por la tarea*. Doktoretza-tesia, Euskal Herriko Unibertsitatea.